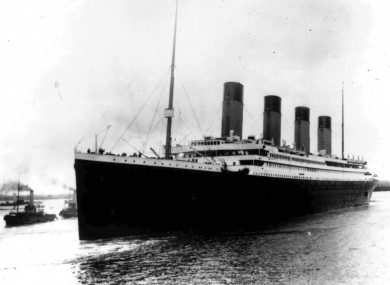
An exploration into what factors affected survival in the sinking of the Titanic



# 

# 

[**Dataset:**](#_lx7wycn4w27w) **2**

[Description:](#_d1fkdek1blqn) 2

[Overview:](#_k64mvtjmi94a) 2

[Collection:](#_55punqp3daa5) 3

[**Data Importation and Cleansing:**](#_sz32mxy2ihwd) **3**

[Importation:](#_6cld3bilrt45) 3

[Cleansing:](#_6zblkza0gixr) 4

[**Exploratory Data Analysis**](#_vxnotcesou0b) **8**

[Heatmap of Correlations:](#_h8tt7b82wh3v) 8

[Survival rates:](#_rfmgi2lq5h8t) 9

[Survival Rates by Gender:](#_efugq35f76kq) 10

[Survival Rates by Passenger Class:](#_8dmv685wulou) 11

[Survival Rates by Fare:](#_lich1jkctw0) 13

[Survival Rates by Location of Embarkation:](#_fvpc06vma3oa) 14

[Survival Rates by Title](#_g0jq77ngcypy) 15

[**Percentage Chances of Survival:**](#_uypg7dsxpk4z) **16**

[Passenger Class:](#_33fwwjvrtpt6) 17

[Title:](#_3w54ih8uzk0q) 17

[Gender:](#_wm6djxah27xj) 18

[**Questions**](#_rjhgxrqmo393) **19**

[Does gender affect chance of survival?](#_z0kl6e5fpbs1) 19

[Is there a difference in the average fare of those who survived and those who died?](#_z0kl6e5fpbs1) 19

[Is there a correlation between passenger class and survival?](#_6rdh2yos0hyh) 19

[Predicting the fare of a passenger aged 45](#_kkwdh44q6ppm) 19

[**Analysis**](#_2y855x7sfhpy) **19**

[Does gender affect chance of survival?](#_t7xd13rxxbsz) 19

[Is there a difference in the average fare of those who survived and those who died?](#_8vg7vra3isac) 23

[Is there a correlation between passenger class and survival?](#_7gp4hkw4rtq8) 24

[Predicting the fare of a passenger aged 45](#_qac9t2ffcrow) 24

[**Conclusion:**](#_xab9geaxuaep) **25**

# Dataset:

## Description:

For this project, I chose to analyse a dataset based on the sinking of the Titanic (<https://en.wikipedia.org/wiki/RMS_Titanic>).The Titanic was a British passenger ship that was built in Belfast, Northern Ireland. There was an estimated 2,224 crew and passengers aboard, of which around 1,500 died. The ship sank after hitting an iceberg around 600km south of Newfoundland.

## Overview:

The dataset has 14 columns:

* Age: Age of passenger
* Cabin:Cabin number
* Fare: Price of ticket
* Embarked: Where the passenger boarded the ship

(C = Cherbourg; Q = Queenstown; S = Southampton)

* Name: Name of passenger
* Parch: Number of parents/children aboard the ship
* PassengerId: ID of passenger
* Pclass: Class of passenger (1st, 2nd, 3rd)
* Sex: Gender of passenger (Male or Female)
* SibSp: Number of Siblings / Spouses aboard the ship
* Survived: If the passenger survived ( 0 = did not survive, 1 = survived)
* Ticket: Ticket number
* Title: Passenger title (Mr, Mrs, etc)
* Family\_Size: combination of SibSp and Parch to get size of family

## Collection:

I got this dataset from Kaggle, it can be found at the following link in csv format: <https://www.kaggle.com/jamesleslie/titanic-cleaned-data#train_clean.csv>

# Data Importation and Cleansing:

## Importation:

I started off by importing pandas

import pandas as pd

I then used pandas to read in the dataset from the csv file and stored it in a pandas dataframe called “df”

df = pd.read\_csv("data/titanic.csv")

I then printed out the dataframe shape to see how many rows and columns were present

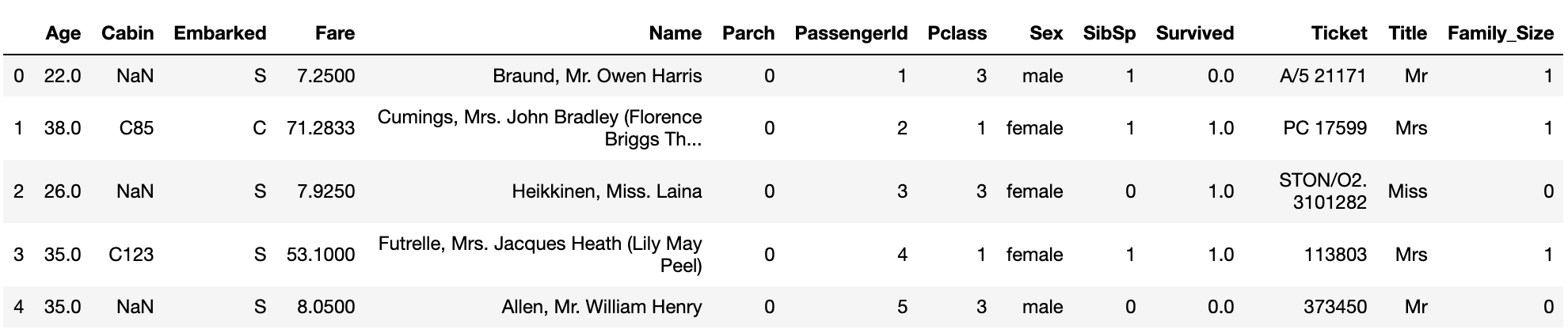
print("dataframe shape" , df.shape)



As we can see, the dataframe has 891 rows and 14 columns.

I then printed out the head of the dataframe to display the top 5 rows

df.head()

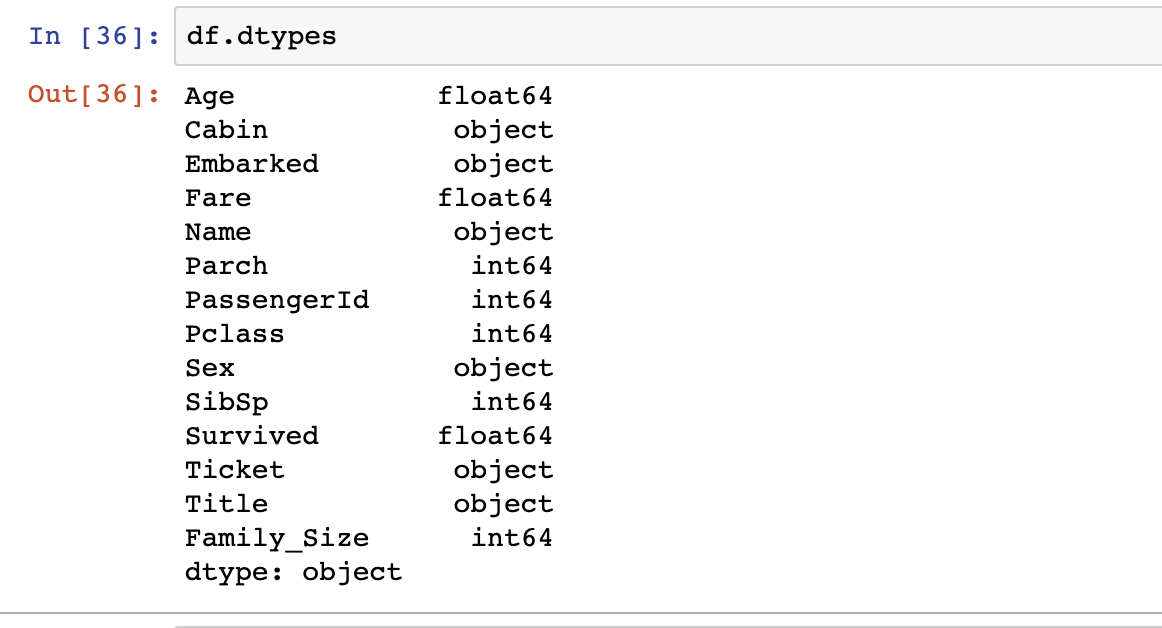


## Cleansing:

To check the data types of the columns, I used the following command:

df.dtypes

This resulted in the following output:



Dealing with each column individually:

**Age:** This column refers to the age of the passengers and hence does not need to be a float, as ages are whole numbers, so this will be converted to an integer.

df.Age = df.Age.astype(int);

**Cabin:** This refers to the cabin number of the passenger, and as there are a limited number of cabins aboard the ship, this data is categorical, and will be converted as such. The NaN values will also be replaced with the value ‘none’.

df['Cabin'] = df['Cabin'].cat.add\_categories('None')

df['Cabin'].fillna('None', inplace =True)

df.Cabin = df.Cabin.astype('category')

**Embarked:** This refers to the location the passenger boarded the ship, and as there are only a select number of possible locations to embark from, this is categorical and shall be converted as such.

df.Embarked = df.Embarked.astype('category')

**Fare:** This refers to the fee paid by the passenger to board the ship, and will remain as a float, as the number needs precision (even if it only needs it to two decimal places).

**Name:** The names of the ships passengers, this will be deleted as it is not necessary for the investigation.

del df['Name']

**Parch:** The number of parents or children of the passenger, this will remain as an integer as there cannot be a fraction of a person aboard the ship.

**PassengerId:** This is the unique ID of the passenger and will remain as an integer as it is an identifying value.

**Pclass:** This refers to the passenger’s class, and will be changed from an integer to category as there are only 3 potential values for class, 1st, 2nd and 3rd class.

df.Pclass = df.Pclass.astype('category')

**Sex:** This refers to the gender of the passengers, and this will be categorical as there are only two possible values, male or female.

df.Sex = df.Sex.astype('category')

**SibSp:** This refers to the number of siblings or spouses a passenger has, and will remain an integer as a person has to be a whole number.

**Survived:** This refers to whether a passenger lived or died, this is categorical data as the potential values are 1 and 0, for survived and died, and will be converted to categorical, being first converted to an int to remove the floating point accuracy.

df.Survived = df.Survived.astype(int);

df.Survived = df.Survived.astype(‘category’);

I originally planned to use categorical on this column, but as I performed my analysis I realised it had to be numerical to perform some calculations

**Ticket:** This refers to the code for the passengers ticket, which will be removed as it is is not necessary for this investigation.

del df['Ticket']

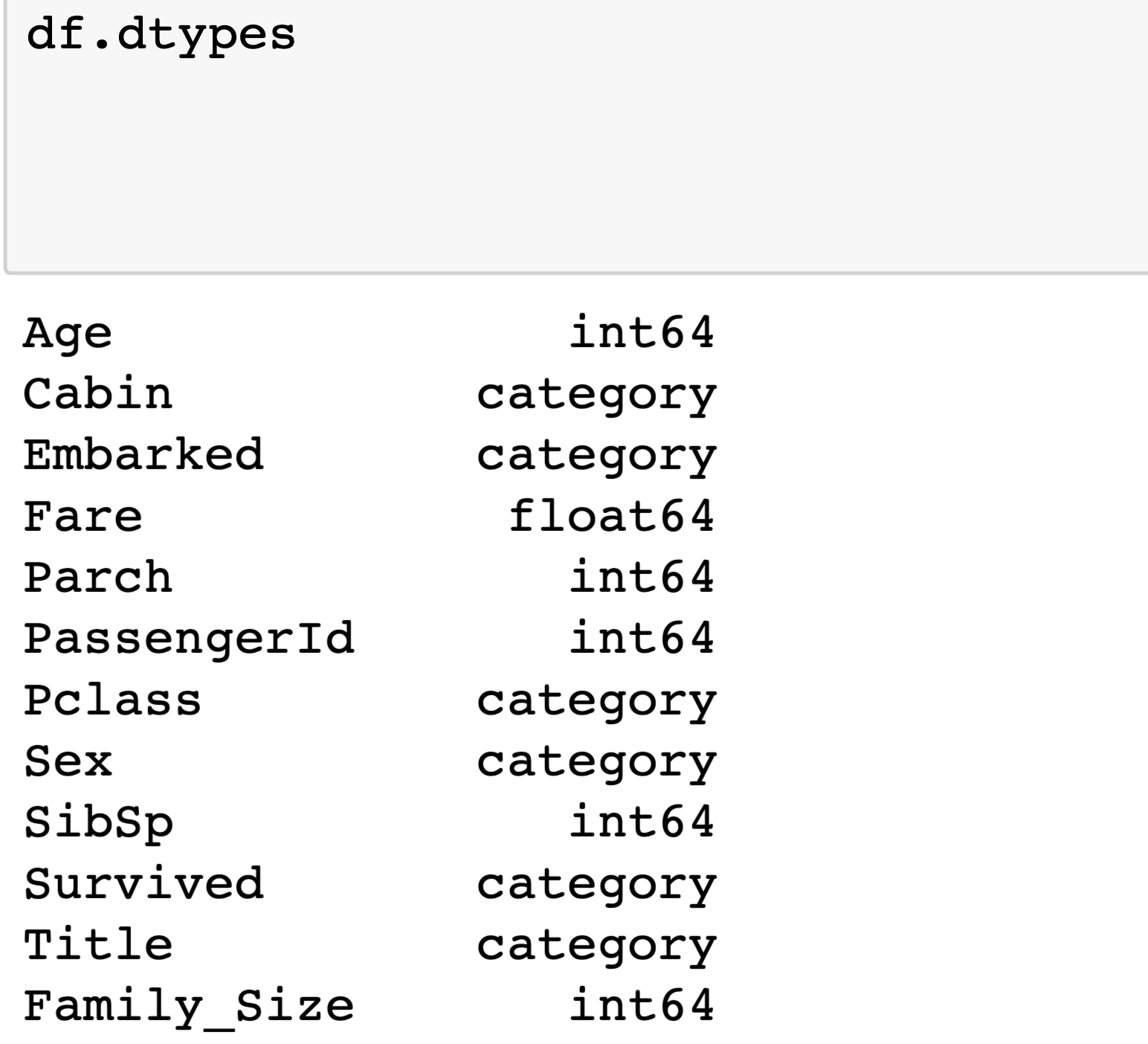
**Title:** This refers to the titles of the passengers aboard the ship, eg Mr, Mrs, etc. This will be converted to categorical as there are only a finite amount of potential values for titles.

df.Title = df.Title.astype('category')

**Family\_Size:** This refers to the size of the family aboard the ship, and will be as an integer as it must be a positive whole number, as we are counting people.

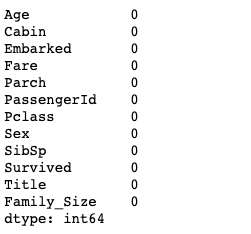
Now, to check the data types again.

df.dtypes



We can see that the data types of the columns have been converted successfully.

df.isna().sum()

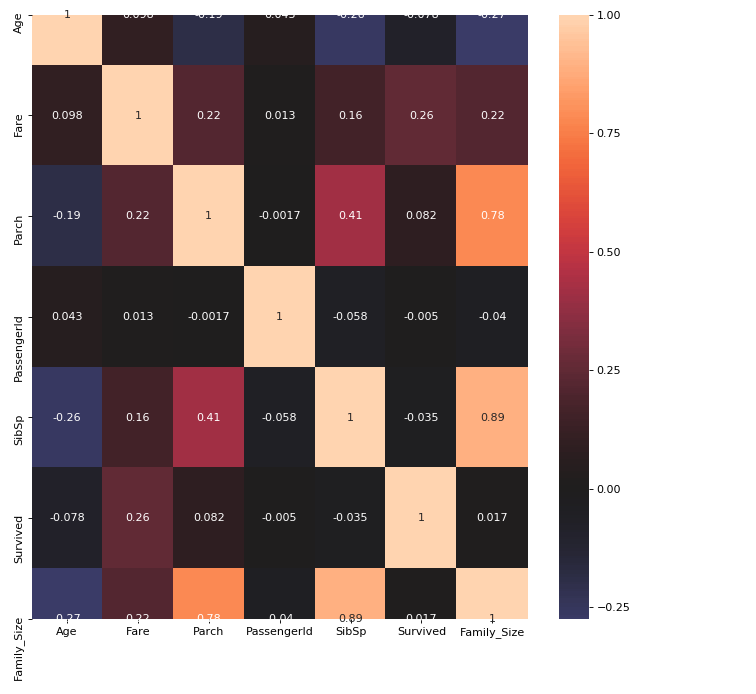


We can see that the dataframe has been cleared of all null values.

# 

# Exploratory Data Analysis

## Heatmap of Correlations:



Interpretation: As expected, there is a strong correlation between Parch and Family\_Size, as well as SibSp and Family\_Size, as family size is made of a combination of these two columns

All other fields appear to have a weak or non existent correlation.

## Survival rates:

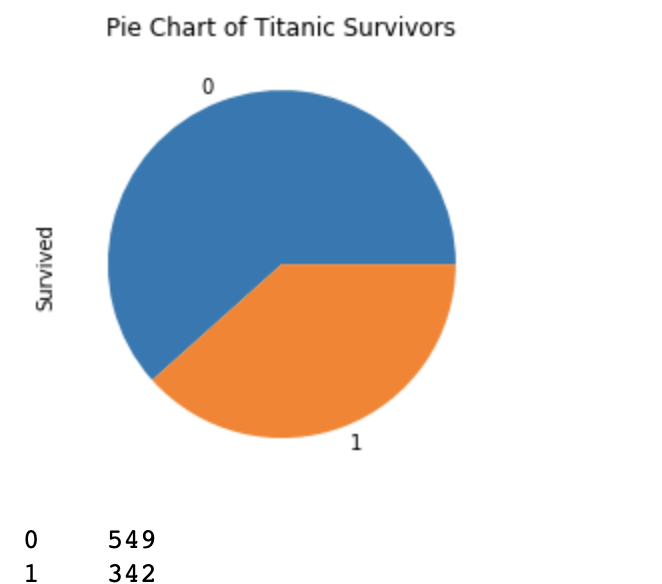
survived = df.Survived.value\_counts()

survived.plot(kind='pie', labels=df.Survived)

plt.title("Pie Chart of Titanic Survivors")

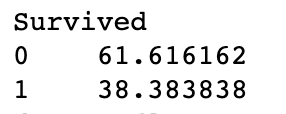
plt.show()

df.Survived.value\_counts()



As we can see, from our sample 549 people died and 342 people survived the shipwreck.

df.groupby("Survived").size()/df.shape[0]\*100

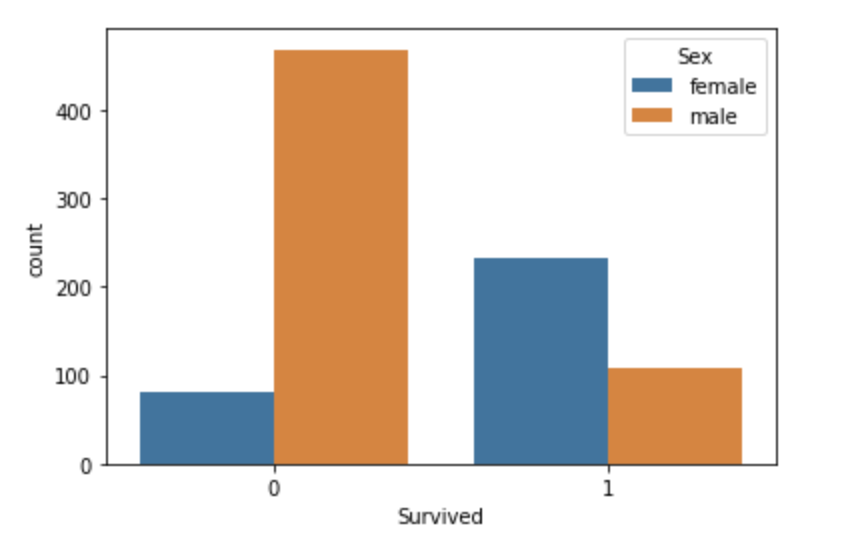


This equates to a 38.38% survival rate.

### Survival Rates by Gender:

import seaborn as sns

sns.countplot(x='Survived', hue="Sex", data=df)



* We can see from this bar chart that a large percentage of those who died were men.
* Of the survivors, there were around twice as many women as there were men

Interpretation: Gender appears to have an impact upon survival rates, as an inspection of the bar chart appears to show that more men died than women.

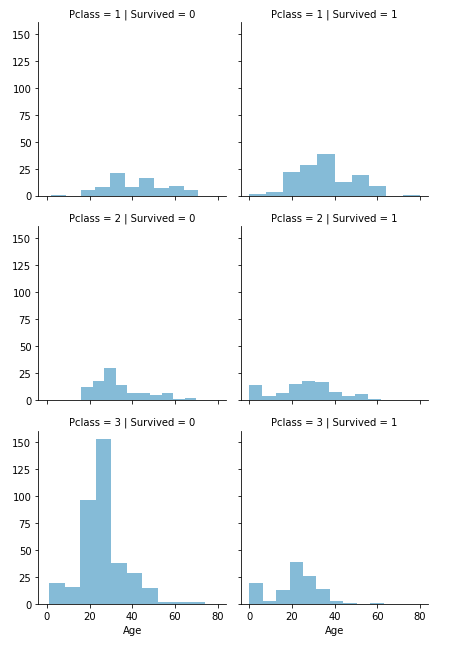
### 

### Survival Rates by Passenger Class:

facetgrid = sns.FacetGrid(df, col='Survived', row='Pclass')

facetgrid.map(plt.hist, 'Age', alpha=.5)

facetgrid.add\_legend();



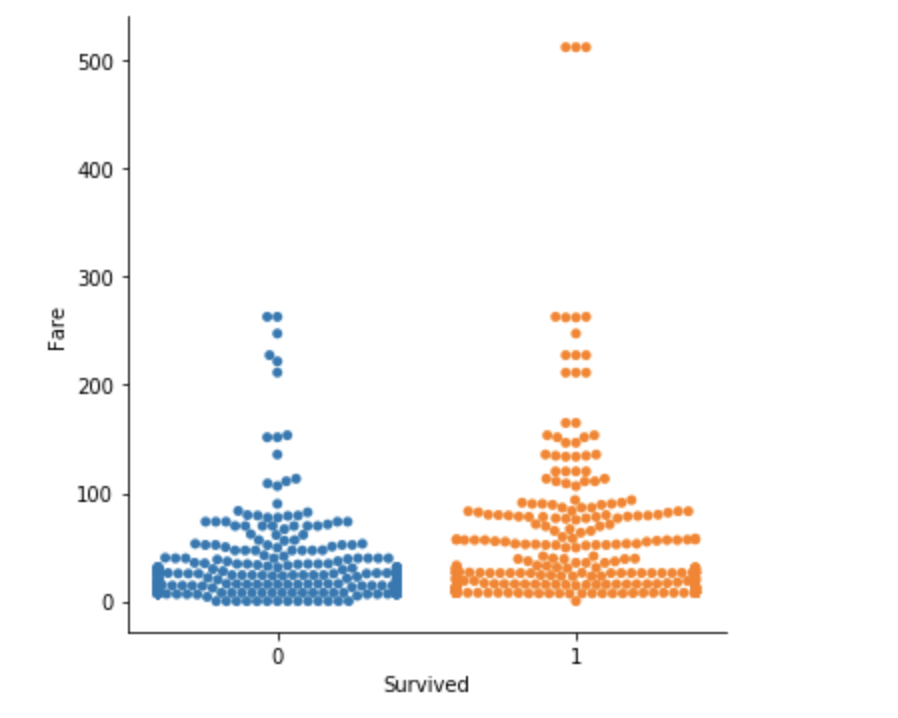
As we can see from the facet grid, per class:

* First class had the largest amount of survivors in its class
* Second class was evenly spread between survived and died
* Third class was least likely to survive, with a large percentage of third class dying

Interpretation: It appears that passenger class may have an impact on survival rate, as people in third class appear to have a poorer survival rate than those in first class.

### Survival Rates by Fare:

sns.catplot("Survived", "Fare", kind="swarm", data=df);



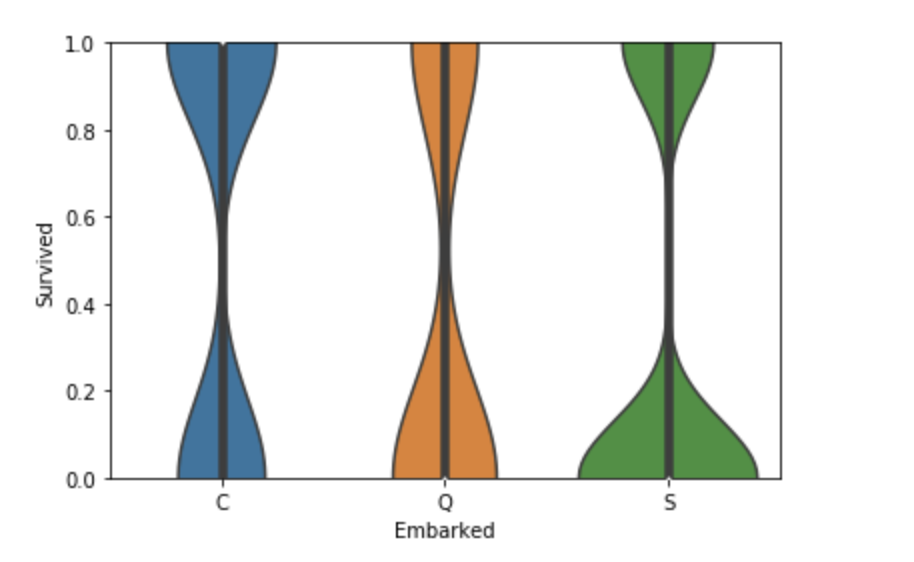
Interpretation: As we can see, there is a fairly equal spread between those who died and didnt based on their fairs. Although, those who spent more than 300 on their fares all survived as we can see from the swarm plot, so we can deduce that the passengers fare did have an impact on their survival rate to a certain extent.

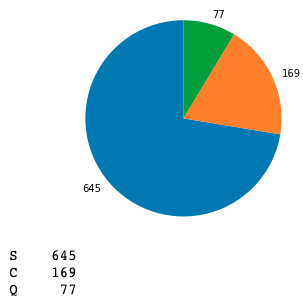
### 

### 

### Survival Rates by Location of Embarkation:

sns.violinplot("Embarked", "Survived", data =df).set\_ylim(0, 1)





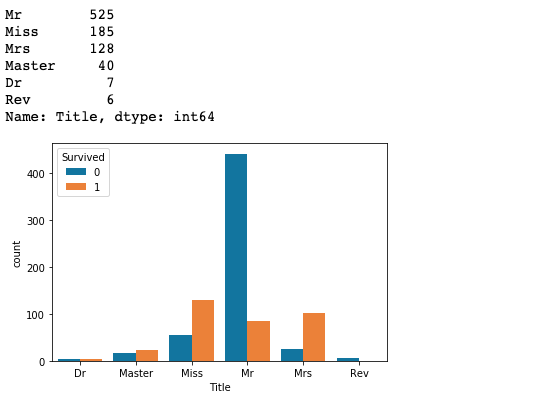
Interpretation: As we can see, there appears to be a correlation between survival and embarkation location, and from a visual inspection this graph is very similar to the Passenger class vs Survival violin plot above, where more people in first class survived than in third.

### 

### Survival Rates by Title

sns.countplot(x='Title', hue="Survived", data=df)

df.Title.value\_counts()



Interpretation:

There were very few doctors, which appear to have a slightly lower chance of survival than death, both being roughly equal.

There were few Masters aboard the titanic, having what appears to be a roughly equal chance of survival.

Of all the Misses aboard, it appears that at least twice as many survived as died.

Of all the titles, Mr was the most common, but also having the third most survivors, showing a low survival rating.

Out of all the passengers aboard, Mrs had the highest survival rating per title, with around 3 - 4 times as many people surviving as dying.

Reverend was the most unfortunate title to hold aboard the titanic, as even though only 6 of them were aboard, all of them perished.

# Percentage Chances of Survival:

Method for getting survival rating percentage:

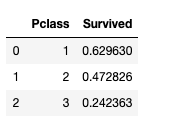
def get\_survival\_rating(x , y ):

return df[[x, y]].groupby([x], as\_index=False).mean().sort\_values(by=y, ascending=False)

### Passenger Class:

survival\_rating\_pclass = get\_survival\_rating('Pclass' , 'Survived')

survival\_rating\_pclass

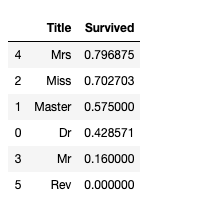


Interpretation: Those in first class had a 63% chance of survival, second class had a 47% chance of survival, and those in 3rd class had a 24% chance of survival. For third class, that’s less than one in four people who survived. This clearly shows that being in a higher class grants you a higher chance of surviving.

### Title:

survival\_rating = get\_survival\_rating('Title' , 'Survived')

survival\_rating



Interpretation:

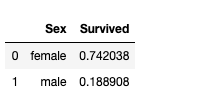
We can see females had the highest survival rates. With Mrs (married women) and Miss (unmarried women) having the highest survival rates of all titles. In third comes Master (unmarried men). Doctors had the 4th highest survival rate with 43%, although this title doesn’t infer a gender, as a male or a female can be a doctor. Mr had the second lowest survival rate, with just 16% surviving of the 525 males as discovered above. Reverend had the lowest survival rate ( 0%).

### 

### Gender:

survival\_rating = get\_survival\_rating('Sex' , 'Survived')

survival\_rating



Interpretation: It is clear that being female gives you a higher percentage chance of surviving the titanic, with 74% of females surviving, compared to 18% of males surviving.

# Questions

## Does gender affect chance of survival?

## Is there a difference in the average fare of those who survived and those who died?

## Is there a correlation between passenger class and survival?

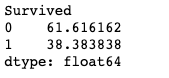
## Predicting the fare of a passenger aged 45

# 

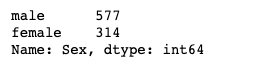
# Analysis

## Does gender affect chance of survival?

df.groupby("Survived").size()/df.shape[0]\*100



df.Sex.value\_counts()



The average survival rate for all of the passengers on the ship as observed above was seen to be 38.38%

The total number of passengers aboard the ship in our sample was 891.

Taking the chance of survival for male and female to be equal we expect the survival outcomes to be the following:

571 males \* .3838 = 219.14

=> 219 males should have lived.

571 - 219 = 352

=> 352 males should have died

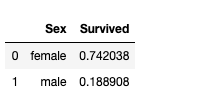
314 \* .3838 = 120.51 = 121

=> 121 females should have lived

314 - 121 = 193

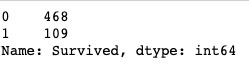
=> 193 females should have died.

Observed survival rates:



male\_df = df.loc[df['Sex'] == 'male']

male\_df.Survived.value\_counts()

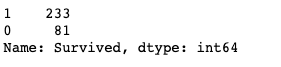


109 men survived

468 men died

male\_df = df.loc[df['Sex'] == 'female']

male\_df.Survived.value\_counts()



233 women survived

81 women died

expected = [121, 193, 219, 352]

obs = [233, 81, 109, 468]

We put these results into arrays in the format female survived, female died, male survived and male died for both expected and observed

We then calculate our degrees of freedom

Since there are two columns and 4 rows

We get degrees of freedom as

(4-1) ( 2-1) = 3

Hypotheses:

* H0: There is no relationship between Sex and Survival
* H1: There is a relationship between Sex and Survival

Confidence interval: α = 0.05

Decision rule: We reject H0 if pObs > p( α)

Calculations:

expected = [121, 193, 219, 352]

obs = [233, 81, 109, 468]

chi2, pvalue, \_, \_ = st.chi2\_contingency([expected, obs])

print('pvalue', pvalue)



p\_alpha = st.chi2.ppf(0.95, 3)

print('p(α)' , p\_alpha)



Decision: We can reject H0 as pObs > p alpha

We can conclude that there is a relationship between sex and survival.

## Is there a difference in the average fare of those who survived and those who died?

survived\_df = df.loc[df['Survived'] == 1]

survived\_fares = survived\_df.Fare

died\_df = df.loc[df['Survived'] == 0]

died\_fares = died\_df.Fare



H0 There is no difference in the average fare of those who survived and those who died

H1: There is a difference in the average fare of those who survived and those who died

Confidence interval: α = 0.05

Decision Rule: We reject H0 if tObs < α

Calculation:

Calculate t test for first 100 people in the data set as a sample

st.ttest\_ind(survived\_fares[0:100],died\_fares[0:100])



Decision: We fail to reject H0 as our P value is 0.12 which is greater than our alpha of 0.05

We do not have enough statistical evidence to prove that there is a difference in the average fare of those who survived and those who died

## Is there a correlation between passenger class and survival?

def get\_pearsonr\_correlation(x , y ):

return st.pearsonr(x, y)

correlation = get\_pearsonr\_correlation(df.Pclass, df.Survived)

correlation



As we can see, there is a moderate negative correlation between passenger class and survival.

## Predicting the fare of a passenger aged 45

slope, intercept, r\_value, p\_value, std\_err = st.linregress(df.Age,df.Fare)

slope, intercept, r\_value, p\_value, std\_err

predicted\_fare = intercept + slope \* 45

print('predicted fare is: ' ,predicted\_fare)

Forumula for linear regression is: Ε(y) = (β0 +β1 x).



We can see that this formula gave us an output of 37.90, which is the predicted fare of a passenger aged 45.

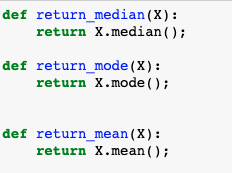
# Conclusion:

From the sample size I used (891) I feel like the conclusions I arrived at were accurate, as the total amount of passengers aboard the Titanic was around 2,400. This means the data set I used was just over a third of the total passengers.

In my exploratory data analysis I discovered some interesting relationships between the different factors that affect survival in the sinking of the Titanic, even if I couldn’t mathematically prove that some of the factors were related, such as passenger class and survival.

On a positive note, I did prove the connection between sex and survival, as women and children were favoured for survival in the accident.

I did make some modules reusable, such as the mean, median and mode, which could be easily reused for anything in python that is numerical to find these values.



I also made the function for getting the pearsonr correlation reusable



I also made a reusable function for survival rating, which is easily reusable within this project



Potential hypothesis errors are Type I and Type II errors that could have happened during my analysis:

A type I error is rejecting the null hypothesis and accepting the alternative hypothesis even though the null hypothesis is true.

A type II error rejecting the alternative hypothesis and accepting the null hypothesis even though the alternative hypothesis is true.